KLASIFIKASI MENGGUNAKAN ALGORITMA DATA MINING

Ferry Susanto[1]

[1]Mahasiswa Program Studi Informatika Universitas Mitra Indonesia

ABSTRAK

Data Mining merupakan Penambangan data dari sekumpulan fakta yang terekam dan di ekstraksi menjadi pengetahuan, klasifikasi data mining diantaranya adalah dengan Algoritma C4.5, ID3, -Nearest Neighbor, Naïve Bayesian Clasification, CART (Clasification And Regression Tree), Pohon Keputusan merupakan cara bagaimana mengklasifikasi Algoritma terhadap data mining dalam hal ini dengan pengklasifikasian yang tepat dapat menghasilkan Pohon Keputusan yang baik.

Keyword : Data Mining, C4.5, Pohon Keputusan

1. Pendahuluan

Metode Estimasi merupakan salah satu metode yang ada dalam Data Mining. Ada hal yang perlu dipahami bahwasanya metode ini dapat bekerja apabila himpunan data sebagai sampel data yang akan di proses bersifat numerik dan memiliki label. Biasanya metode ini tidak memiliki rumus yang pasti karena bersifat Regresi. Artinya dalam penentuan sebuah keputusan dari sebuah sampel baru berasal dari sebuah rumus yang terbentuk berdasarkan parameter-parameter himpunan data.

Dalam metode estimasi terdapat beberapa algoritma yang dapat dijadikan sebagai Learning Algorithma diantaranya yaitu Regresi Linier. Klasifikasi merupakan sebuah proses training (pembelajaran) suatu fungsi tujuan (target) yang digunakan untuk memetakan tiap himpunan atribut suatu objek ke satu dari label kelas tertentu yang di definisikan sebelumnya. Teknik Klasifikasi ini cocok digunakan dialam mendeskripsikan data-set dengan tipe data dari suatu himpunan data yaitu biner atau nominal. Adapun kekurangan dari teknik ini yaitu tidak tepat untuk himpunan data ordinal karena pendekatan-pendekatan yang digunakan secara implisit dalam kategori data.

Ada beberapa teknik klasifikasi yang digunakan sebagai solusi pemecahan kasus diantaranya yaitu:

- Algoritma C4.5

- Algoritma K-Nearest Neighbor
- ID3
- Naïve Bayesian Clasification
- CART (Clasification And Regression Tree)

Dan lain-lain

Output atau keluaran dari metode klasifikasi ini biasanya dalam bentu “Decision Tree (pohon keputusan)”. Dalam pembahasan kali ini saya mencoba untuk membahas tentang Algoritma C4.5.

2. Pembahasan

2.1 Algoritma C4.5

Algoritma C4.5 merupakan salah satu solusi pemecahan kasus yang sering digunakan dalam pemecahan masalah pada teknik klasifikasi. Keluaran dari algoritma C4.5 itu berupa sebuah decision tree layaknya teknik klasifikasi lain. Sebuah pohon keputusan adalah sebuah struktur yang dapat digunakan untuk membagi kumpulan data yang besar menjadi himpunan-himpunan record yang lebih kecil dengan menerapkan serangkaian aturan keputusan. Dengan masing-masing rangkaian pembagian, anggota himpunan hasil menjadi mirip satu dengan yang lain (Berry & Linoff, 2004).

Adapun penjelasan tentang Algoritma C4.5 itu sendiri yaitu Salah satu algoritma C4.5 induksi pohon keputusan yaitu ID3 (Iterative Dichotomiser 3). input berupa sampel training, label training dan atribut. Algoritma C4.5 merupakan pengembangan dari ID3.

Jika suatu set data mempunyai beberapa pengamatan dengan missing value yaitu record dengan beberapa nilai variable tidak ada, jika jumlah pengamatan terbatas maka atribut dengan missing value dapat diganti dengan nilai rata-rata dari variable yang bersangkutan. (Santosa, 2007) Untuk penyelesaian kasus didalam Algoritma C4.5 ada beberapa elemen yang diketahui yaitu: 1. Entropy 2. Gain Entropy(S) merupakan jumlah bit yang diperkirakan dibutuhkan untuk dapat mengekstrak suatu kelas (+ atau -) dari sejumlah data acak pada ruang sampel S. Entropy dapat dikatakan sebagai kebutuhan bit untuk menyatakan suatu kelas. semakin kecil nilai

Entropy maka akan semakin Entropy digunakan dalam mengekstrak suatu kelas.

Entropi digunakan untuk mengukur ketidakaslian S. Adapun rumus untuk mencari nilai
Entropi.

$$Entropy(S) \equiv -p_{\oplus} \log_2 p_{\oplus} - p_{\ominus} \log_2 p_{\ominus}$$

Dimana:

S : ruang (data) sampel yang digunakan untuk pelatihan

$p_{\oplus}$: jumlah yang bersolusi positif atau mendukung pada data sampel untuk kriteria tertentu

$p_{\ominus}$: jumlah yang bersolusi negatif atau tidak mendukung pada data sampel untuk kriteria tertentu.

- Entropi(S) = 0, jika semua contoh pada S berada dalam kelas yang sama.
- Entropi(S) = 1, jika jumlah contoh positif dan negative dalam S adalah sama.
- 0 > Entropi(S) > 1, jika jumlah contoh positif dan negative dalam S tidak sama.

Gain (S,A) merupakan Perolehan informasi dari atribut A relative terhadap output data S. Perolehan informasi didapat dari output data atau variabel dependent S yang dikelompokkan berdasarkan atribut A, dinotasikan dengan gain (S,A). Adapun rumus untuk mencari nilai Gain yaitu :

$$Gain(S,A) \equiv Entropy(S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy(Si)$$

Dimana:

- A : Atribut
- S : Sampel
- n : Jumlah partisis himpunan atribut A
- |Si| : Jumlah sampel pada pertisi ke –i
- |S| : Jumlah sampel dalam S

Adapun langkah-langkah untuk penyelesaian Algoritma C4.5 terlihat pada siklus di bawah ini :

Algoritma Penyelesaian Algoritma C4.5

2.2 Contoh Kasus dan Penyelesaian

Masalah yang akan di analisis adalah untuk mengklasifikasikan calon pendaftar di suatu STMIK xxx dalam hal pemilihan program studi khususnya : Sistem Komputer Atau Sistem Informasi. Adapun data yang digunakan dalam membentuk pohon keputusan untuk menganalisis minat calon mahasiswa baru untuk mendaftar ke STMIK xxx berdasarkan program studi strata 1 adalah nama mahasiswa, minat calon mahasiswa, asal sekolah, jenis kelamin, hobi. Data selajutnya akan dilakukan pra-proses untuk menghasikan data kasus yang siap dibentuk untuk menjadi sebuah pohon keputusan.

Data yang tidak lengkap disebabkan karena ada data yang kosong atau atribut yang salah. Demikian pula dengan data minat calon mahasiswa baru yang mendaftar ke STMIK xxx berdasarkan program studi strata 1, ada sebagian atribut yag tidak perlu sehingga proses Data Preprocessing perlu dilakukan sehingga data base sesuai dengan ketentuan yang diperlukan.

Data Preprocessing merupakan hal yang penting dalam proses data mining, hal yang termasuk antara lain:

1. Data Selection

Data minat calon mahasiswa/i baru yang mendaftar ke STMIK xxx berdasarkan program studi strata 1 tersebut akan menjadi data kasus dalam proses operasional data mining. Dari data yang ada, kolom yang diambil sebagai atribut keputusan adalah hasil, sedangkan kolom yang diambil atribut penentuan dalam pembentukan pohon keputusan adalah:

a. Nama Mahasiswa
b. Minat calon mahasiswa
c. Asal sekolah
d. Jenis kelamin
e. Hobi

2. Data Preprocessing / Data Cleaning

Data Cleaning diterapkan untuk menambahkan isi atribut yang hilang atau kosong dan merubah data yang tidak konsisten.

3. Data Transformation

Dalam proses ini, data ditransferkan ke dalam bentuk yang sesuai untuk proses data mining.

4. Data Reduction

Reduksi data dilakukan dengan menghilangkan atribut yang tidak diperlukan sehingga ukuran dari database menjadi kecil dan hanya menyertakan atribut yang diperlukan dalam proses data mining, karena akan lebih efisien terhadap data yang lebih kecil.

Masalah klasifikasi berakhir dengan dihasilkan sebuah pengetahuan yang dipresentasikan dalam bentuk diagramyang biasa disebut pohon keputusan (decision tree). Data berikut ini dipergunakan untuk data latihan. Data selengkapnya tampak pada tabel dibawah ini:

No.	Nama Mahasiswa	Calon Mahasiswa	Asal Sekolah	Kelamin	Hobi	Hasil
1	Novita Devi Batu Bara	Hardware	SMK Komputer	Laki-Laki	Non	SK
2	Ahmad Riyandi	Hardware	SMK Komputer	Laki-Laki	IT	SK
3	Reza Adriansyah	Umum	SMK Komputer	Laki-Laki	Non	SI
4	Gafar Dwi Satrio	Software	SMA UMUM	Laki-Laki	Non	SI
5	Nur Azizah Dalimunthe	Software	SMK TEKNIK	Perempuan	Non	SI
6	Roy Ishak Permana Barus	Software	SMK TEKNIK	Perempuan	IT	SI
7	Muhammad Rizky Fadly	Umum	SMK TEKNIK	Perempuan	IT	SI
8	Zulfikar Ali	Hardware	SMA UMUM	Laki-Laki	Non	SK
9	Putra Mustaqim	Hardware	SMK TEKNIK	Perempuan	Non	SI
10	Debby Latifah Simatupang	Software	SMA UMUM	Perempuan	Non	SI
11	Daniel Alberto Sihombing	Hardware	SMA UMUM	Perempuan	IT	SI
12	Asri Anzani Br. Tarigan	Umum	SMA UMUM	Laki-Laki	IT	SI
13	Abdul Alim	Umum	SMK Komputer	Perempuan	Non	SI
14	Akbar Widiantara	Software	SMA UMUM	Laki-Laki	IT	SK

Tabel Sampel yang digunakan

Keterangan :

Untuk Asal Sekolah yang disebut SMK Komputer yaitu yang berasal dari jurusan Teknik Komputer Dan Jaringan, Multimedia, dan Rekayasa perangkat lunak sedangkan yang dikatakan sekolah umum yaitu Sekolah Menengah Atas yang terdiri dari jurusan IPA maupun IPS dan yang dimaksud SMK Teknik adalah yang berasal dari jurusan baik Teknik Elektro, Teknik Mesin, Teknik Listrik dan Lain-lain. SI merupakan Nilai Atribut Hasil Sistem Informasi dan SK merupakan Nilai Atribut Hasil Sistem Komputer.

Setelah kita memperoleh data Minat Calon Mahasiswa/i Baru yang tercantum pada Tabel Sampel. Langkah selanjutnya adalah menentukan nilai Entropy dan Gainnya:

1. Nilai Entropy
   a. Entropy Total=$Entropy(S) = \sum_{i=l}^{n} - \ pi{*}log_2 \, pi$
      Entropy Total=((-4/14*log 2 (4/14) + (-10/14*log 2 (10/14))
      = 0.863120569

   b. Entropy Minat Calon Mahasiswa
      - Nilai atribut “Hardware” = ((-3/5)*Log 2(3/5)+(-2/5)*log 2 (2/5)
        =0.970950594
      - Nilai atribut “Software” = ((-1/5)*Log 2(1/5)+(-4/5)*log 2 (4/5)
        = 0.721928095
      - Nilai atribut “Umum” = ((-0/4)*Log 2(0/4)+(- 4/4)*log 2 (4/4)
        = 0
   c. Entropy Histori Pendidikan (Asal Sekolah)
      - Nilai atribut “SMK Komputer”
        = ((-2/4)*Log 2(2/4)+(-2/4)*log 2 (2/4) = 1
      - Nilai atribut “SMK Teknik”
        = ((-0/4)*Log 2(0/4)+(- 4/4)*log 2 (4/4) = 0
      - Nilai atribut “SMA Umum”
        =((-2/6)*Log 2(2/6)+(- 4/6)*log 2 (4/6) = 0.918295834
   d. Entropy Hobi
      - Nilai atribut “IT”
        = ((-4/6)*Log 2(4/6)+(- 2/6)*log 2 (2/6) = 0.918295834
      - Nilai atribut “Non IT”
        = ((-2/8)*Log 2(2/8)+(-6/8)*log 2 (6/8) = 0.811278124
   e. Entropy Jenis Kelamin
      - Nilai atribut “1” = ((-4/7)*Log 2(4/7)+(- 3/7)*log 2 (3/7)
        = 0.985228136
      - Nilai atribut “0” = ((-0/7)*Log 2(0/7)+(-7/7)*log 2 (7/7)

= 0

2. Nilai Gain

   Berikut ini adalah nilai Gain dari setiap kriteria.

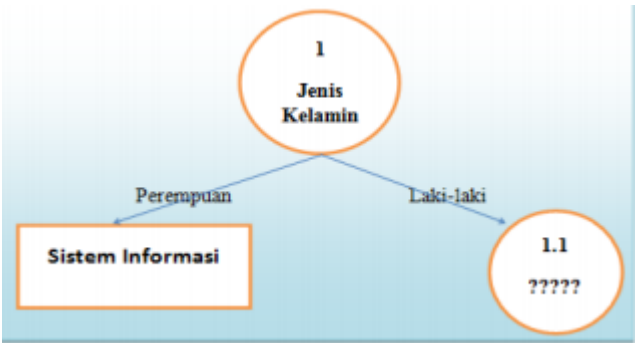
   - Nilai Gain Minat Calon Mahasiswa
     = 0.863120569-((5/14)* 0.970950594))+((5/14)* 0.721928095))+((4/14)*0)))
     = 0.258521037
   - Nilai Gain Histori Pendidikan
     = 0.863120569-((4/14)*1))+((4/14)*0))+((6/14)* 0.918295834)))
     = 0.183850925
   - Nilai Gain Hobi
     = 0.863120569–((6/14)* 0.918295834))+((8/14)*0))
     = 0.005977711
   - Nilai Gain Jenis Kelamin
     = 0.863120569–((7/14)* 0.985228136))+((7/14)*0))
     = 0.005977711

Setelah di dapatkan nilai Entropy dan Gain dari sampel data yang dimiliki, berikut ini adalah rekapitulasi perhitungan nilai Entropy dan Gainnya.

.

NODE		Keterangan	Jml Kasus (S)	Sistem Komputer (SK)	Sistem Informasi (SI)	Entropy	Gain
1	TOTAL		14	4	10	0.86312	
	Minat Calon						0.25852103
		Hardware	5	3	2	0.970950	
		Software	5	1	4	0.721928	
		Umum	4	0	4	0	
	Asal Sekolah						0.18385092
		SMK Komputer	4	2	2	1	
		SMK Teknik	4	0	4	0	
		SMA	6	2	4	0.918295	
	Jenis Kelamin						0.37050650
		Laki-laki	7	4	3	0.985228	
		Perempuan	7	0	7	0	
	Hobi						0.00597771
		IT	6	4	2	0.918295	
		Non	8	2	6	0.811278	

Rekapitulasi Hasil

Tabel di atas menunjukkan bahwasanya kriteria Jenis Kelamin memiliki nilai Gain yang paling tinggi. Untuk fase selanjutnya adalah pembentukan Tree (pohon keputusannya). Berikut ini adalah Tree dari rekapitulasi nilai Entropy dan Gainnya :

Node

Pohon keputusan di atas belum terlihat keputusan yang dominan dari setiap program studi yang di pilih. Maka kita harus mencari kembali nilai Entropy dan Gain dari setiap atribut(kritera) Jenis Kelamin = Laki-laki. 1. Nilai Entropy Berikut ini adalah tabel penyelesainnya. Tabel: Sampel Data Yang Di Uji Ulang (Kriteria Jenis Kelamin)

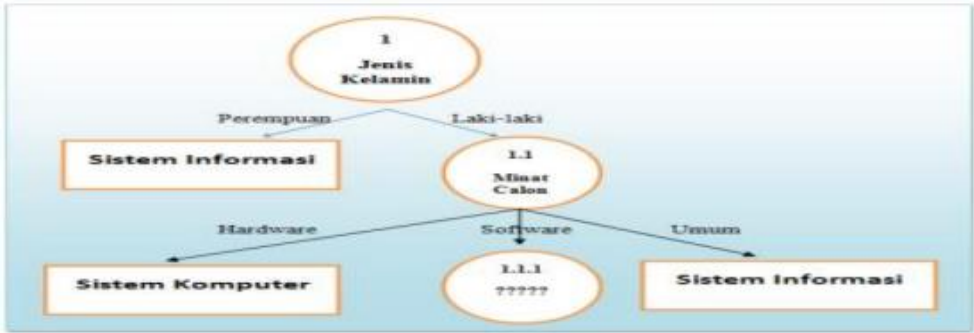
Kriteria	Attribut	Jumlah Kasus	SI	SK
Jenis Kelamin	**Laki-laki**	**7**	**4**	**3**

Setelah itu kita hitung nilai Entropy dari atribut Jenis Kelamin = Laki-Laki yang memiliki jumlah kasus "7" seperti terlihat pada Tabel di bawah ini.

No.	Nama Mahasiswa	Minat Calon Mahasiswa	Asal Sekolah	Jenis Kelamin	Hobi	Hasil
1	Novita Devi Batu Bara	Hardware	SMK Komputer	Laki-Laki	Non	SK
2	Ahmad Riyandi	Hardware	SMK Komputer	Laki-Laki	IT	SK
3	Reza Adriansyah	Umum	SMK Komputer	Laki-Laki	Non	SI
4	Gafar Dwi Satrio	Software	SMA UMUM	Laki-Laki	Non	SI
8	Zulfikar Ali	Hardware	SMA UMUM	Laki-Laki	Non	SK
12	Asri Anzani Br. Tarigan	Umum	SMA UMUM	Laki-Laki	IT	SI
14	Akbar Widiantara	Software	SMA UMUM	Laki-Laki	IT	SK

Langkah selanjutnya adalah menghitung nilainya, Tabel berikutnya menunjukkan hasil Rekapitulasi nilai Entropi dan Gainnya.

NODE		Keterangan	Jml Kasus (S)	Sistem Komputer (SK)	Sistem Informasi (SI)	Entropy	Gain
1.1	Jenis Kelamin = Laki-laki		7	4	3	0.98522	
	Minat Calon						0.69951
		Hardware	3	3	0	0	
		Software	2	1	1	1	
		Umum	2	0	2	0	
	Asal Sekolah						0.02024
		SMK Komputer	3	2	1	0.91829	
		SMK Teknik	0	0	0	0	
		SMA Umum	4	2	2	1	
	Hobi						0.02024
		IT	3	2	1	0.91829	
		Non	4	2	2	1	

Berdasarkan tabel di atas terlihat bahwasanya Attribut = Minat Calon memiliki nilai Gain Tertinggi, maka untuk Root selanjutnya pada pohon keputusannya dapat terlihat pada gambar pohon (tree) berikut ini :

Pohon Keputusan

Karena pohon keputusan belum terlihat keseluruhan hasilnya sehingga kita perlu untuk mencari kembali Nilai Gain dan Entropy selanjutnya berikut ini adalah tabelnya.

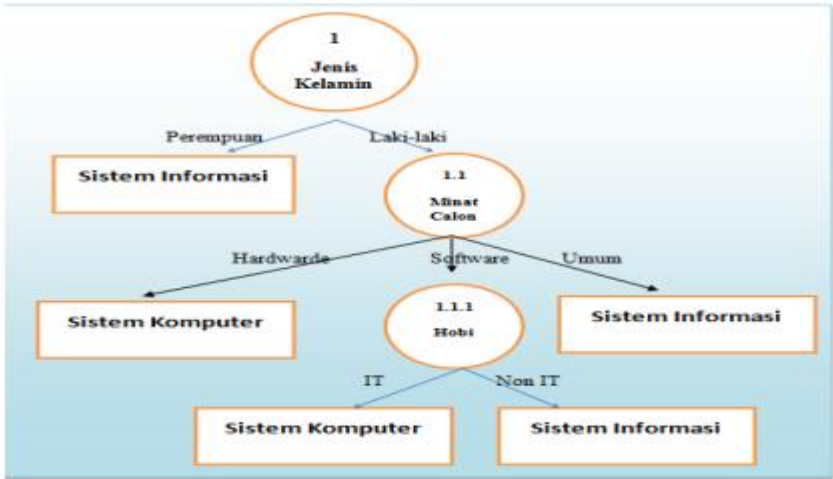
Kriteria	Attribut	Jumlah Kasus	SK	SI
Minat Calon	Software	2	1	1

Setelah itu kita data terlebih dahulu dari atribut Minat Calon = Software yang memiliki jumlah kasus "2" seperti terlihat pada Tabel di bawah ini.

No.	Nama Mahasiswa	Minat Calon	Asal Sekolah	Hobi	Hasil
1	Gafar Dwi Satrio	Software	SMA UMUM	Non	SI
2	Akbar Widiantara	Software	SMA UMUM	IT	SK

Selanjutnya adalah kita menghitung kembali nilai Entropy dan Gainnya seperti terlihat pada tabel di bawah ini:

NODE		Keterangan	Jml Kasus (S)	Sistem Komputer (SK)	Sistem Informasi (SI)	Entropy	Gain
1.1.1	Jenis Kelamin = Laki-laki Dan Minat Calon= Software		2	1	1	1	
	Asal Sekolah						0
		SMK Komputer	0	0	0	0	
		SMK Teknik	0	0	0	0	
		SMA Umum	2	1	1	1	
	Hobi						1
		IT	1	1	0	0	
		Non IT	1	0	1	0	

Gambar di atas menjelaskan bahwasanya yang memiliki kriteria memiliki nilai Gain tertinggi yaitu : 1 maka node pohon keputusannya adalah sebagai berikut:

Hasil dari Pohon Keputusan

3. Penutup

3.1 Kesimpulan

Maka basis pengetahuan atau rule yang terbentuk yaitu :

1. Jika Jenis Kelamin = Perempuan maka Hasil = Sistem Informasi
2. Jika Jenis Kelamin = Laki-laki dan Minat Calon = Hardware maka Hasil = Sistem Komputer
3. Jika Jenis Kelamin = Laki-laki dan Minat Calon=Umum maka Hasil = Sistem Informasi
4. Jika Jenis Kelamin = Laki-laki dan Minat Calon= Software dan Hobi = IT maka Hasil = Sistem Komputer
5. Jika Jenis Kelamin = Laki-laki dan Minat Calon= Software dan Hobi=Non IT maka Hasil = Sistem Informasi

Dari data diatas maka penulis menarik kesimpulan bahwa hasil yang didapat adalah peminat untuk Bidang Sistem Komputer lebih banyak dibanding dengan Bidang Sistem Informasi.

3.2 Saran

Adapun saran-saran yang disampaikan berdasarkan hasil pengamatan dan analisa selama melakukan penelitian adalah: 1. Penelitian selanjutnya sebaiknya menggunakan data yang lebih banyak agar menghasilkan rules yang lebih akurat. 2. Penelitian selanjutnya sebaiknya menggunakan atribut yang lebih banyak agar menghasilkan data yang lebih akurat.

DAFTAR PUSTAKA

1. Pratama, I. W., & Hafiz, A. (2019). Implementasi Data Mining Untuk Menentukan Trend Penjualan Cetakan Sablon Pada Fatih Clothing Di Bandar Lampung. *Jurnal Cendikia*, *18*(1), 326-329.
2. Turban , Efraim & Aronson, Jay E. 2001. Decision Support Systems and Intelligent Systems. 6th edition. Prentice Hall: Upper Saddle River, NJ
2. L. A. Abdillah, "Students learning center strategy based on e-learning and blogs," in Seminar Nasional Sains dan Teknologi (SNST) ke-4 Tahun 2013, Fakultas Teknik Universitas Wahid Hasyim Semarang 2013, pp. F.3.15-20.
3. V Wiratna Sujarweni, Metodologi Penelitian. Yogyakarta: PUSTAKABARUPERSS,2014.
4. Kusrini and Emha Taufiq Luthfi, Algoritma Data Mining. Yogyakarta: ANDI, 2009.
5. Rahmadya T. H dan Herlawati Prabowo P. W, Penerapan Data Mining dengan Matlab. Bandung: Rekayasa Sains, 2013.
6. Hermawati. F. Astuti. (2013). Data Mining. Yogyakarta: Andi Offset.